Replication: An Alternative Softmax Operator for Reinforcement Learning

Author names
Department of Computer Science
National Yang Ming Chiao Tung University
{xxx, yyy, zzz}@nycu.edu.tw

1 Problem Overview

Please provide a brief overview of the selected paper. You may want to discuss the following aspects:

- The main research problem tackled by the paper
- High-level description of the proposed method

2 Background and The Algorithm

Please present the essential background knowledge and the algorithm in this section. You may also describe the notations and the optimization problem of interest.

3 Detailed Implementation

Please explain your implementation in detail. You may do this with the help of pseudo code or a figure of system architecture. Please also highlight which parts of the algorithm lead to the most difficulty in your implementation.

3.1 Lunar Lander Domain

Experiment Settings

The paper's settings:

- network: a hidden layer comprised of 16 units with RELU activation functions + a second layer with 16 units and softmax activation functions
- use REINFORCE to train the network
- batch episode size: 10
- learning rate = 0.005
- optimizer: Adam
- do 10 experiments
  - Boltzmann softmax: $\beta = 1, 2, 3, 5, 10$
  - Mellowmax: $\omega = 3, 5, 7, 8, 11$
- each experiments do 400 runs, each runs train 40000 episodes

Our expiriment settings are basically the same as the paper's, the only different is we do each experiment 3 runs (seed=22, 321, 7654), each run trains 15000 episodes. Because the training needs a lot of time, so we simplified this part.

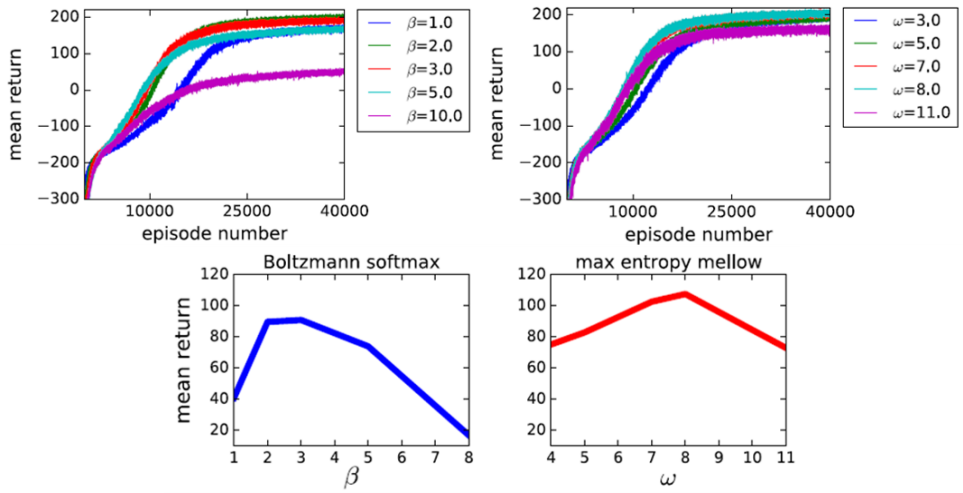
4 Empirical Evaluation

Please showcase your empirical results in this section. Please clearly specify which sets of experiments of the original paper are considered in your report. Please also report the corresponding hyperparameters of each experiment.

4.1 Lunar Lander Domain

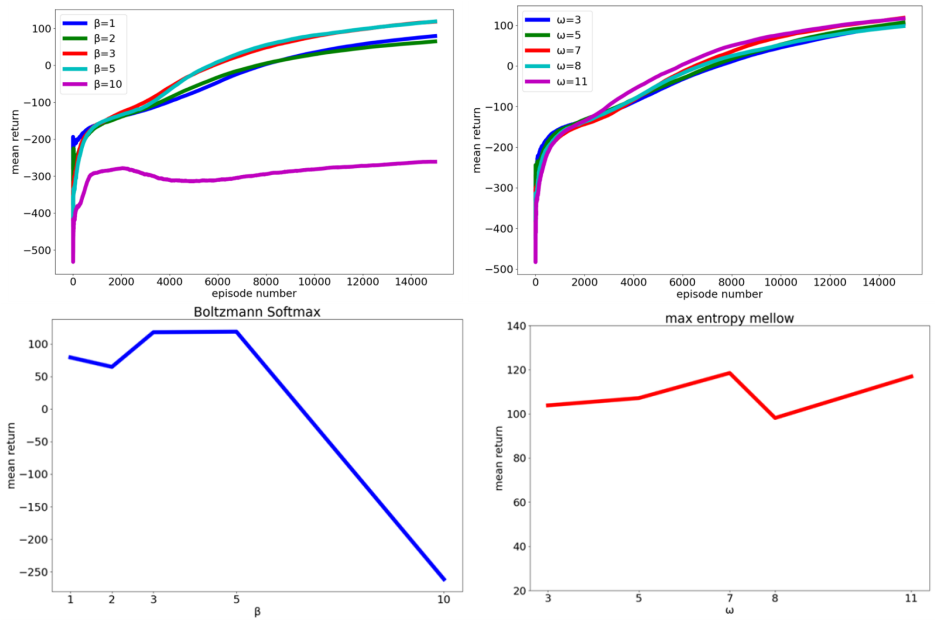
Experiment Results

The paper's results:



In the paper's results, the Boltzmann softsmax method performs well when $\beta = 2, 3$, and the Mellowmax method performs well when $\omega = 7, 8$.The Boltzmann softmax method, $\beta = 10$ performs the worst in the 10 experiments, it gets a much lower average mean return than others.The Mellowmax method, $\omega = 8$ performs the best in the 10 experiments.

Our results:



In our results, the Boltzmann softmax method performs well when $\beta = 3, 5$, and the Mellowmax method performs well when $\omega = 7, 11$. The Boltzmann softmax method, $\beta = 10$ performs the worst in the 10 experiments and gets a much lower average mean return than others. The Mellowmax method, $\omega = 7$ performs the best in our 10 experiments.

Comparison

Our experiment results are a little different from the paper's, and we think it is because that we use much less runs and less episodes in each run. The mean return of Boltzmann softmax $\beta = 10$ is about -300, and it's much lower than the paper's result. We find that it gets a lot of negative rewards at seed=22, its ewma reward is less than -1000 during almost all episodes, however, its ewma reward reaches 200 while seed=7654. It seems that Boltzmann softmax $\beta = 10$ is unstable, so the chosen seed affects a lots, and more runs are needed to get precise results.

5 Conclusion

Please provide succinct concluding remarks for your report. You may discuss the following aspects:

- The potential future research directions
- Any technical limitations: The Mellowmax method needs more time to solve beta, In LunarLander, it takes about 0.0014 sec per step with Boltzmann softmax and 0.0046 sec per step with Mellomax.
- Any latest results on the problem of interest

References